

LIBRARY OF THE UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

510.84 IlGr ho.61-80 cop.3



Digitized by the Internet Archive in 2013



UNIVERSITY OF ILLINOIS GRADUATE COLLEGE DIGITAL COMPUTER LABORATORY

REPORT NO. 74

GENERATING PSEUDO-RANDOM NUMBERS IN THE ILLIAC

Вy

Carl C. Farrington, Jr.

December 7, 1956

This work has been supported by Contract N6ori-71 Task XXX United States Navy ONR NR 044 001



GENERATING PSEUDO-RANDOM NUMBERS IN THE ILLIAC

Carl C. Farrington, Jr., University of Illinois

Introduction

This report contains results of research directed toward finding a better method of generating pseudo-random numbers in the ILLIAC. When we speak, loosely, of random numbers we mean numbers generated by a random process. We shall call a process random if we cannot predict the next number to be generated.

Since very long sequences of random numbers are required in Monte Carlo computations, reading such numbers into the computer is not practical. A possible alternative would be to generate inside the computer numbers by a random physical process. This approach has the disadvantages that additional equipment is required and that the computations are not reproducible. The other alternative is to generate numbers by a determinate arithmetic process. Of course, such numbers will not be "random." However, inasmuch as one is only interested in obtaining trustworthy answers to problems, such "pseudorandom" numbers may be satisfactory. Of course, whether the results of a computation for which a particular sequence of pseudo-random numbers is used should be given as much confidence as those for which a randomly selected sequence of numbers is used, depends upon the particular problem. Accordingly, the user must decide for himself how much confidence to place in the numerical results he obtains.

These considerations led to placing primary emphasis on obtaining a fast method of generating numbers, one taking only about 1/2 millisecond per number, which generates a very long sequence. This is the sense in which we speak of a "better" method.

For the most part this report contains the results of three tests on two methods of generating pseudo-random numbers. These tests each determine whether some particular property of the numbers is statistically predictable. One of the methods tested is that used in the present library routine. The other is a modification of a method proposed by Töcher (1).

⁽¹⁾ Töcher, K. D., "The application of automatic computers to sampling experiments," J. Roy. Stat. Soc., ser. B, vol. 16, (1954) 39-61.



I. Methods of generating pseudo-random numbers

The method used in the present library routine is the "middle-of-the-square" method (2). A certain 38 binary digit number $(2^{-19} + 3 \times 2^{-38})$ proposed by N. Metropolis is taken as the starting point. This number is squared and the middle 38 digits of the product are used as the second member of the sequence. This is squared to obtain the third member, and so on. The first 1,000 numbers are usually discarded. The sequence is about 700,000 numbers long, after which it rapidly degenerates to a sequence of all zeros.

The first objective of this research was to find a faster method, one taking on the order of one-half millisecond to generate one number. Accordingly, processes involving the use of multiplication or division orders had to be ruled out. The method finally arrived at uses five starting numbers. A linear combination of the last four is added to a number obtained by permuting the digits of the first number, to obtain the 6th number of the sequence. The same linear combination of the 3rd, 4th, 5th and 6th numbers is then added to the permuted digits of the 2nd number to obtain the 7th number, and so on. The linear combination used has only integral coefficients so the linear combination can be formed by a series of additions, subtractions, and shifts. The shifting operations are used simultaneously to permute the digits of the other number. Secondly, the red tape is reduced by generating five numbers instead of one each time the routine is entered. The overall result is that the routine takes 3 milliseconds to generate five numbers, so that the numbers are generated at a rate of 0.6 millisecond per number.

The method proposed by Töcher was to form a linear combination of a fixed number of the preceding numbers to obtain the next number in the sequence, without permuting the digits of any of them. He recommended that the sequence of binary digits composing this sequence of numbers be paired and that these pairs be replaced by single digits according to the following code.

00 reject 01 0 10 1 11 reject

⁽²⁾ Hammer, Preston C., "The Mid-Square Method of Generating Digits," Proceedings of a Symposium July 29, 30 and July 1, 1949, National Bureau of Standards Applied Mathematics Series 12.



He demonstrated that this scheme would improve several types of non-randomness in the original sequence. He recommended that the resulting sequence of digits be regrouped as variables and subjected to the pairing process again to obtain the final sequence of random numbers.

Since the ILLIAC is a parallel machine, the time which would be consumed in coding digits in the manner described would be absolutely prohibitive for obtaining long sequences of numbers.

It was thought that a linear combination might be used directly as a method of generating pseudo-random numbers without the use of the improvement scheme suggested. One sequence, the Fibonacci numbers, which involves a linear combination of only two of the preceding numbers is rather well known. The Fibonacci numbers are obtained by making each number the sum of the two preceding numbers, using 0, 1 as the starting numbers, so that one obtains 0, 1, 1, 2, 3, 5, ... The starting numbers are represented in the computer as 0×2^{-39} and 1×2^{-39} . The sequence increases quite rapidly, and integers larger than 2^{40} are soon obtained. Overflow occurs and the sum must be regarded as a sum modulo 2^{40} . The Fibonacci numbers and other sequences of numbers obtained as linear combinations of preceding numbers have been found to have a periodicity of their terminal digits which makes them unsuitable as sources of random numbers. An analysis showing this result follows.*

We consider a sequence of integers modulo 240 given by

$$A_{n+q} = \sum_{k=1}^{q} a_k A_{n+q-k} , n \ge 0$$
 (1)

with starting numbers A_0 , A_1 , ..., A_{q-1} . We assume that the last coefficient a_q is an odd integer. This is a sufficient condition for periodicity of the terminal digits, for if a_q is odd it is relatively prime to any power of 2. Hence for any s, a_q has an inverse modulo 2^s , so that the equation

$$a_{q}^{A} = A_{n+q} - \sum_{k=1}^{q-1} a_{k}^{A} + q - k$$

has a unique solution modulo 2^s for A_n in terms of its first q successors. It is clear that if the s terminal digits of q successive numbers in the

^{*} The results and proof given here were suggested by those in an article by Vern Hoggatt, "A type of periodicity for Fibonacci numbers," Mathematics Magazine, vol. 28, no. 3, (1955), 139-142, analyzing the Fibonacci numbers in their decimal representation.



sequence are the same as the s terminal digits of q successive preceding numbers

$$T_s(A_{n+i+M}) = T_s(A_{n+i})$$
, i = 0, 1, ..., q - 1 (2)

where $T_s(A)$ represents the s terminal binary digits of A, then the s terminal digits will be periodic for $i \ge 0$ of period M. Since there are only a finite number of possibilities for the s terminal digits of any number, the existence of such numbers n and M is assured. But if the coefficient a_q is odd, then equation (2) will hold for all $i \ge -n$ due to the uniqueness of the predecessor of any q numbers. Thus the periodicity of the terminal digits will begin at the very beginning of the sequence. This result can be stated as (R) Let A_k be a sequence of integers modulo 2^{40} given by (1) with a_q odd. Then for any integer $s \le 40$ there exists an integer M such that

$$T_s(A_k + M) = T_s(A_k)$$

for all k.

If we denote the period of the s terminal binary digits by $\mathbf{p}_{_{\mathbf{S}}}$ we can show now that

$$p_{s} \le 2^{s-1}p_{1}$$

To obtain this result we consider an auxiliary sequence

$$F_{n+q} = \sum_{k=1}^{q} a_k F_{n+q-k}, n \ge 0;$$

 $F_{0} = F_{1} = \dots = F_{q-2} = 0; F_{q-1} = 1.$

(The Fibonacci numbers, for which q = 2 and $a_1 = a_2 = 1$ are a special case of this) We shall show by induction that

$$A_{n+q+p} = \sum_{k=1}^{q} \sum_{j=0}^{p} a_{j+k} F_{p+q-1-j} A_{n+q-h}, n, p \ge 0$$
 (3)

using the convention that $a_k = 0$ for k > q. For p = 0 we have for the right hand member

$$(\sum_{k=1}^{q} a_k^A_{n+q-k}) F_{q-1} = A_{n+q}$$

by equation (1), since $F_{q-1} = 1$, which holds for all $n \ge 0$. The assumption that equation (3) holds for p = r and for all $n \ge 0$ yields

$$A_{n+q+(r+1)} = A_{(n+1)} + q + r$$

$$= \sum_{k=1}^{q} \sum_{j=0}^{r} a_{j+k} F_{r+q-1-j} A_{n+1+q-k}$$

$$= \sum_{k=1}^{q} \sum_{j=0}^{r} a_{j+1+k-1} F_{r+q-(j+1)} A_{n+q-(k-1)}$$

$$= \sum_{k=0}^{q-1} \sum_{j=1}^{r+1} a_{j+k} F_{r+q-j} A_{n+q-k}$$

Using the convention that $a_k = 0$ for k > q we may write

If r + 1 < q the second summation may be written

$$\sum_{j=1}^{r+1} a_j F_{r+q-j} A_{n+q} = \sum_{j=1}^{q} a_j F_{r+q-j} A_{n+q} - \sum_{j=r+2}^{q} a_j F_{r+q-j} A_{n+q}$$

$$= \sum_{j=1}^{q} a_j F_{r+q-j} A_{n+q}$$

for $F_r = \dots = F_{q-2} = 0$. If $r + 1 \ge q$, since $a_j = 0$ for j > q, we again obtain



From the defining equation we have

$$\sum_{j=1}^{r+1} a_{j} F_{r+q-j}^{A} A_{n+q} = (\sum_{j=1}^{q} a_{j} F_{r+q-j}) A_{n+q} = F_{r+q}^{A} A_{n+q}$$

From equation (1) the third summation is

$$\left(\sum_{k=1}^{q} A_k A_{n+q-k}\right) F_{r+q} = A_{n+q} F_{r+q}$$

so we obtain

$$A_{n+q+(r+1)} = \sum_{k=1}^{q} \sum_{j=0}^{r+1} a_{j+k} F_{r+q-j} A_{n+q-k} + F_{r+q} A_{n+q} - A_{n+q} F_{r+q}
 = \sum_{k=1}^{q} \sum_{j=0}^{r+1} a_{j+k} F_{(r+1)+q-1-j} A_{n+q-k}$$

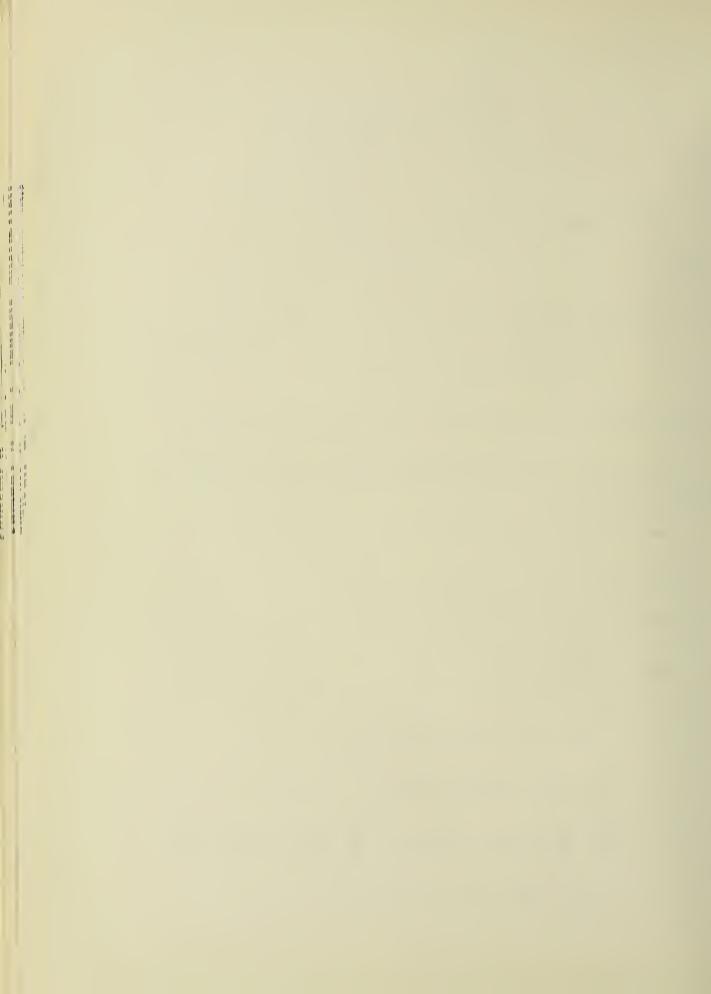
which establishes the result for p = r+1. Hence it is true for all p, n ≥ 0 .

If M is the period of the s terminal digits of the F's, then there are integers B, such that

$$F_{M+1} = B_1 2^s$$
, i = 0, 1, ..., q - 2
 $F_{M+q-1} = B_{q-1} 2^s + 1$

Hence, since M must be greater than q - 1, equation (3) implies

$$\begin{split} F_{2M+i} &= \sum_{k=1}^{q} \sum_{j=0}^{q-1} a_{j+k} \ F_{M+q-l-j} \ F_{M+i-k} \\ &= \sum_{j=0}^{q-1} \sum_{k=1}^{q} a_{j+k} \ F_{M+q-l-j} \ F_{M+i-k} \\ &= \sum_{j=0}^{q-1} \sum_{t=j+1}^{q} a_{t} \ F_{M+q-l-j} \ F_{M+i+j-t} \\ &= \sum_{j=0}^{q-l-i} \sum_{t=1}^{q} a_{t} \ F_{M+i+j-t} \ F_{M+q-l-j} - \sum_{j=1}^{q-l-i} \sum_{t=1}^{j} a_{t} \ F_{M+i+j-t} \ F_{M+q-l-j} \\ &+ \sum_{j=q-i}^{q-l} \sum_{t=j+1}^{q} a_{t} \ F_{M+i+j-t} \ F_{M+q-l-j} \end{split}$$



$$= \sum_{j=0}^{q-1-i} F_{M+i+j} F_{M+q-1-j} + \sum_{j=q-i}^{q-1} \sum_{t=j+1}^{q} a_t F_{M+i+j-t} F_{M+q-1-j}$$

$$-\sum_{j=1}^{q-1-i} \sum_{t=1}^{j} a_t F_{M+i+j-t} F_{M+q-1-j}$$

$$(4)$$

We note that for i < q - 1 the first summation is

$$\underbrace{\sum_{j=0}^{q-1-i}}_{F_{M+i+j}} F_{M+q-1-j} = 2F_{M+i} F_{M+q-1} + \underbrace{\sum_{j=1}^{q-2}}_{F_{M+i+j}} F_{M+q-1-j}$$

Since each factor in the summation on the right is a multiple of 2^s , the summation is a multiple of $2^{2s} \ge 2^{s+1}$. The first term on the right is

$$2F_{M+i} F_{M+q-1} = 2B_i 2^s (B_{q-1} 2^{s+1}) = B_i 2^{s+1} + B_i B_{q-1} 2^{2s+1}$$

Except for i = 0 when the second summation in equation (4) is not present, all factors in the second and third summations in equation (4) are multiples of 2^S so that these summations are multiples of $2^{S} > 2^{S+1}$. Hence we have expressions of the form

$$F_{2M+i} = C_i 2^{s+1}$$
 for $i = 0, 1, ..., q-2$

For i = q - l the third summation in equation (4) is not present, the second is a multiple of 2^{s+l} and the first is

$$F_{M+q-1}^2 = B_{q-1}^2 = B_{q-1}^2 + B_{q-1}^2 + 1$$

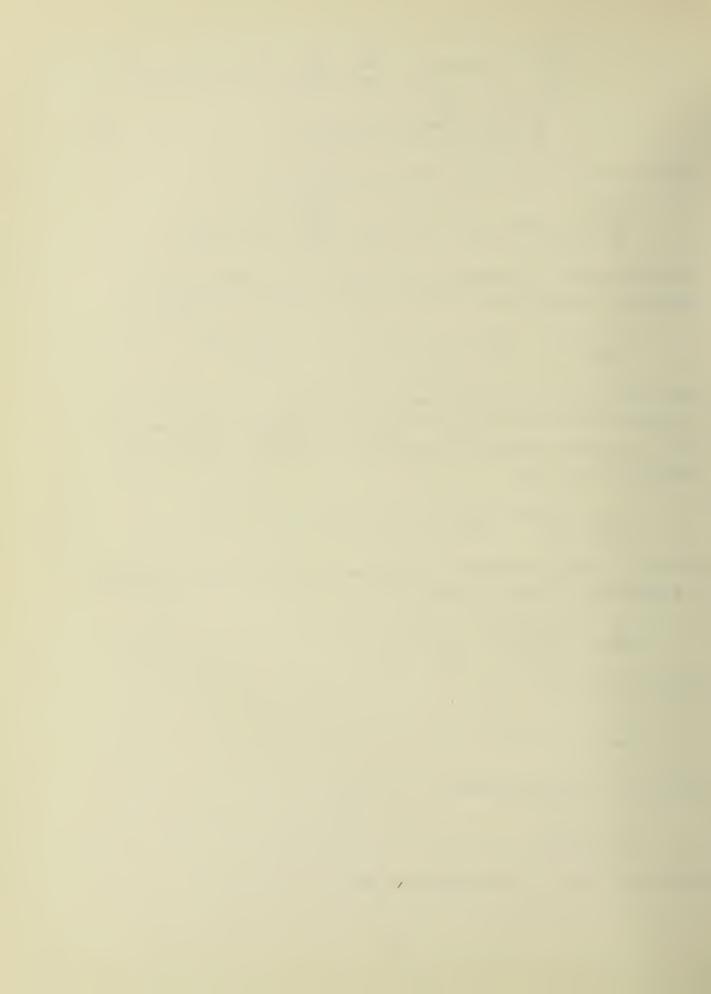
which is of the form

$$F_{2M+q-1} = C_{q-1}2^{s+1} + 1.$$

Hence for the auxiliary sequence

$$p_{s+1} \leq 2p_s$$

and we notice that if m is a number such that



$$F_{p_m} + q-1 = B_{q-1}2^m + 1, B_{q-1} \text{ odd}$$

then

$$p_s = 2^{s-m}p_m$$
, $s \ge m$

Finally, we obtain our result by letting M be the period of the s terminal digits of the F's. Then by equation (3) with n+q=i and p=M

$$A_{M+i} = \sum_{k=1}^{q} \sum_{j=0}^{q-1} a_{j+k} F_{M+q-l-j} A_{i-k}$$

$$= \sum_{j=0}^{q-1} \sum_{t=j+1}^{q} a_{t} A_{i+j-t} F_{M+q-l-j}$$

$$= F_{M+q-l} A_{i} + \sum_{j=1}^{q-1} \sum_{t=1}^{q} a_{t} A_{i+j-t} F_{M+q-l-j} - \sum_{j=1}^{q-1} \sum_{t=1}^{j} a_{t} A_{i+j-t} F_{M+q-l-j}$$

By our previous results this is of the form

$$A_{M+1} = (B_{q-1}2^{s} + 1) A_{1} + D2^{s}$$

Hence

$$T_s(A_{M+i}) = T_s(A_i)$$
, $i \ge 0$

Thus the p_s for the A's must be a divisor of the p_s for the F's for each s. The periods for the A's need not be the same as for the F's. For example, if the starting numbers for the A's are all even numbers then $p_l = 1$. Consequently our result should read

$$p_s(A) \leq 2^{s-1} p_1 (F)$$

The periodicity just demonstrated of the terminal digits of the numbers obtained by the linear combination method makes these digits too predictable to be regarded as random. For this reason the method was modified to include permutation of the digits of one of the numbers of the



combination. If $(a_0, a_1, \dots, a_{39})$ denotes a forty binary digit number the permutation operation \mathcal{T} chosen is given by

$$\pi$$
 (a₀, a₁, ..., a₃₉) = (a₀, a₄, a₅, ..., a₃₉, a₁, a₂, a₃)

With this modification of the linear combination method the condition for periodicity of the s < 39 terminal digits is no longer met, for the s terminal digits of a number no longer depend only on the s terminal digits of the q preceding numbers. The combination investigated uses q = 5 and is given by

$$A_{n+5} = 7A_{n+4} + A_{n+3} - 4A_{n+2} + 3A_{n+1} + \pi A_n$$

This combination has the advantage that the equation

$$\pi A_n = A_{n+5} - 7A_{n+4} - A_{n+3} + 4A_{n+2} - 3A_{n+1}$$

has a unique solution for a member of the sequence in terms of its 5 immediate successors, since the permutation operation \mathcal{T} has a well-defined inverse. The choice of the other coefficients and of q were rather arbitrary, except that one case of a choice of q = 2 was found to be unsatisfactory.

In the machine the number A_{n+5} is formed by first placing A_{n+4} in the A register and A_n in the Q register. A left shift of one place is performed to give 2 A_{n+4} in the A register. To this is added - A_{n+2} and + A_{n+1} giving $2A_{n+4} - A_{n+2} + A_{n+1}$. Then a left shift of two places is performed, and to the result is added - $A_{n+4} + A_{n+3} - A_{n+1}$ to give $4(2A_{n+4} - A_{n+2} + A_{n+1}) - A_{n+4} + A_{n+3} - A_{n+1} = 7A_{n+4} + A_{n+3} + 3A_{n+1}$. In the meantime the digits q_1 , q_2 , q_3 of q_1 have been added to the end of the number in the A register, and the digits q_1 , q_2 , ..., q_3 have been shifted left three places, the last three places in the Q register being replaced by zeros. Adding the number in the Q register to the number in the A register then completes the formation of A_{n+5} .



II. Results of tests

We describe next the tests for randomness performed on the two methods of generating random numbers, the "middle-of-the-square" method and the modified linear combination method, and the results of these tests. In the main three tests were performed. The first was a chi-squared frequency test. The binary digits of the numbers were grouped in fours, each group of four being regarded as a sexadecimal digit. The test was a test of the frequencies of the sexadecimal digits. For the middle-of-the-square method, since 38 digit numbers are produced, the first 36 binary digits of the first number were used as the first 9 sexadecimal digits; the last two binary digits of the first number and the first two of the second number were used as the tenth sexadecimal digit, and the remaining 36 binary digits of the second number were used as the next nine sexadecimal digits. All tests of the middle-of-the-square method began with the 1001st number of the sequence. Since the modified linear combination produces numbers 40 binary digits in length, 10 sexadecimal digits were obtained from each such number.

For the first 4,800,000 sexadecimal digits, representing 3/4 of the sequence, for the middle-of-the-square method the frequency test gave a value of 16.6. The probability of getting a higher value of chi-squared is 35%. For the first 4,800,000 sexadecimal digits obtained by the modified linear combination method a chi-squared value of 12.8 was obtained, which corresponds to a probability of 62%.

The second test performed on the two sequences was a poker test. In this test a "hand" of five sexadecimal digits is classified as a bust, one pair, two pairs, three of a kind, full house, straight, four of a kind, or five of a kind. Hands were tested in groups of 10,000. Testing every 5th group of 10,000 beginning with the first such group making a total of 280,000 hands a chi-squared value of 13.81 was obtained for the middle-of-the-square method. This value corresponds to a probability of 5.6% of obtaining a higher value of chi-squared. For 280,000 hands selected in the same manner a chi-squared value of 7.46 was obtained for the sequence generated by the modified linear combination method. A chi-squared value of 7.46 for seven degrees of freedom corresponds to a probability of 39%.



The third test applied is similar to a run or gap test for columns of binary digits and made use of the logical product instruction of the computer. The logical product of two numbers is formed and tested to determine whether the resulting number contains all zeros. If the probability of any digit of a 40 digit number being zero is one-half, the probability of the logical product of two numbers containing all zeros is $(3/4)^{40}$. If the logical product of the first two numbers did not contain all zeros, the product of it with the third number was formed. If this number was not all zeros its product with the fourth number was formed, and so on. The numbers were divided into groups of 20 and a determination was made of how many factors in the logical product of these numbers were required before a number containing all zeros was obtained. The number of factors were classified as two or less, three, four, ..., twelve, thirteen or more. The number of groups of 20 falling into each of these classifications was compared to the theoretical distribution. The theoretical distribution was slightly different for the 38 digit numbers generated by the middle-of-the-square method than for the 40 digit numbers generated by the modified linear combination method, of course. For the first 700,000 numbers of the sequence obtained by the middle-of-the-square method a chi-squared value of 9.1 was obtained, which corresponds to a probability of 60%. For the first 1,000,000 numbers of the modified linear combination sequence chi-squared was 5.8, for which the corresponding probability is 89%.

This third test has the advantage that it is very fast, since operations are performed on whole numbers rather than on sexadecimal digits. For example, testing the first 1,000,000 numbers of the sequence obtained by the modified linear combination method required less than 30 minutes of computing time. For this reason this test was the first applied to two other trial methods. One of these was a modified linear combination method which used only two starting numbers. This was tried because it was felt that a routine which generated only one number each time it was entered would be more convenient, and this feature would not be too costly in time if only two numbers were used in the combination. However, the chi-squared value obtained for the one sequence was 190.3, which is far beyond the acceptable range.



The method of analysis applied to the linear combination method is not applicable to the modified linear combination method, because the permutation operation \mathcal{T} does not have the necessary algebraic properties. Indeed, there does not appear to be any promise that any method of analysis will yield illuminating results. But perhaps this is not an altogether bad feature of a method for generating random numbers. At any rate, in lieu of an analytic determination of the length of the sequence the first ten million numbers of the sequence were tested for repetition. It was found that the sequence does not repeat within the first ten million numbers. As this test was being made every 100,000th group of five numbers of the sequence was printed out, as an aid to future users of the routine who might wish to check for possible machine error or wish to use different parts of the sequence. This list is given in the appendix.



| 0 | 3500000 | 7000000 |
|--|--|--|
| F085N08L+- J25291+706 63F95019F7 1L6-479F+3 662487-L56 | 9-2F917FN3 4N-JJ23L72 5L0F2J8JJ+ -8N8FFF5N1 3J7068NL-9 | 9+F87-LL86 6-F-N15F9L 342+57+9NJ 2F3-F68N+5 417719F9N4 |
| 500000 | 4000000 | 7500000 |
| 7350338N5+ J108-4N+5- L9N0L5456L -L8N0J41N5 77L-77027J | FOO-80L4+J OLL58L6177 F5604564J7 43-2660+L3 90NL-0N49+ | J928+NN+JF 6191261+J3 1J1N3412+L 068369FN-F -32F9F203J |
| 1000000 | 4500000 | 8000000 |
| -2515+6-3F 674J-946-7 148+21JJ5+ 005-1LN5+8 4956J3918J | 40JJ3J3+F- +NF5F73779 F5F3888+87 64+7-921+1 N-JNF7613F | N57-L7-37N 35364J-N09 F7J129F6FL F5+N3993J8 42704-+916 |
| 1500000 | 500000 | 8500000 |
| -J13L36L51 60360L269- LJJ1082-F0 37473-LJ24 L99F6NL-79 | NNLJJ263N1 8477-6J-5N LNN7960L82 NO5N19J272 OL13FN97F3 | 9N3LOFNON4 4+-5FL8950 21L+2L8LF4 1+0L14J277 F-N61715+1 |
| 2000000 | 5500000 | 9000000 |
| 45FNON57L9 6208J2N18- N+N-2L2N77 L2NJ3+F0J2 4J-F7F8898 | F9F-913F+5 353N4N931N 08JL60J287 3519595J+7 -1471NF0J6 | 6411+668N8 F535273553 -495N5L987 519+N9+N70 09081N-9F5 |
| 2500000 | 6000000 | 9500000 |
| 5JJ4+L9331 L3-4340+F+ O-6498L3N3 +NLL34LLJN +65024+N2L | 6L864FL065 103N688482 01J60-6L25 -608J1+12J 74+454469L | +25F73+L+3 2J-1F4J3+9 9NF6719F45 J34106NN21 915777+34- |
| 3000000 | 6500000 | 10000000 |
| 292+805182 87F39N8L31 8838F1F37+ 3181F7JJ44 NF698N+042 | 7982J-230- 4-3569F605 F3488-J030 -2219FF741 80LJ39-J57 | L94-L3N79+ -+59895367 9L2+6L+3L7 2385719LF4 52-7596J44 |









